LLAMA-2 大语言模型的数学形式

何沧平
cangping@staff.weibo.com
微博

摘　要

LLAMA 是最近几个月最流行的开源大语言模型，本文给出该模型的数学形式。

关键词： LLAMA、大语言模型

Mathematical Principles of LLAMA-2 Model*

He Cangping
cangping@staff.weibo.com
WEIBO.COM

Abstract

LLAMA is the most popular open-source Large Language Model(LLM) model in the last few months. This paper presents its mathematic formulas in detail.

Keywords: LLAMA, Large Language Model, LLM

1　引言

随着 chatGPT 的发布，大语言模型成为人工智能领域的研究热点。LLAMA[3] 开源且性能指标接近 chatGPT，很多机构开发大模型时都以 LLAMA 为基础，例如 Baichuan-7B [1] 采用了与 LLAMA 相同的模型结构。2023 年 7 月 18 日，LLAMA-2[2] 发布，模型架构不变，优化代码性能，同时允许商用。

为了迅速应用于业务、严谨地理论研究，本文给出 LLAMA 模型的数学形式，将程序代码改写为数学公式。程序代码与原论文不一致的地方，以程序代码为准。

---

*完稿日期：2023 年 7 月 31 日

[1] https://github.com/baichuan-inc/baichuan-7B

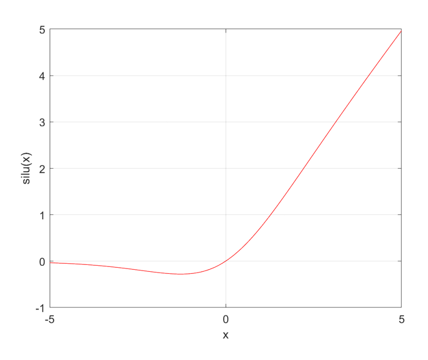


图 1: 激活函数 silu

2 函数定义

作为准备，本节定义几个函数。目前 pytorch 代码中数组的组织方式是行优先，序号从 0 开始，因此本文中的向量、矩阵也按行优先来定义，矩阵元素的序号也从 0 开始。

任意给定正整数 m 和 n，行向量用黑体小写字母表示，形式为 $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n-1})$。矩阵用大写字母表示，形式为

$$X = \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0,n-1} \\ x_{10} & x_{11} & \cdots & x_{1,n-1} \\ \vdots & \vdots & & \vdots \\ x_{m-1,0} & x_{m-1,1} & \cdots & x_{m-1,n-1} \end{bmatrix}.$$

软大函数 (softmax) 定义为

$$\begin{aligned} \mathrm{smax}(\boldsymbol{x}) &= \frac{1}{\sum_{i=0}^{i=n-1} e^{x_i}} (e^{x_0}, e^{x_1}, \ldots, e^{x_{n-1}}), \\ \mathrm{smax}(X) &= \begin{bmatrix} \mathrm{smax}(x_{0:}) \\ \mathrm{smax}(x_{1:}) \\ \vdots \\ \mathrm{smax}(x_{m-1:}) \end{bmatrix} = (\mathrm{smax}(x_{0:}); \mathrm{smax}(x_{1:}); \ldots; \mathrm{smax}(x_{m-1:})), \end{aligned}$$

这里的 $x_{i:} = (x_{i0}, x_{i1}, \ldots, x_{i,n-1})$，圆括号里的分号表示换行。

对向量或矩阵求对数时，对数作用到它们的每一个元素上，即

$$\begin{aligned} \log(\boldsymbol{x}) &= (\log(x_0), \log(x_1), \ldots, \log(x_{n-1})), \\ \log(X) &= \begin{bmatrix} \log(x_{00}) & \log(x_{01}) & \cdots & \log(x_{0,n-1}) \\ \log(x_{10}) & \log(x_{11}) & \cdots & \log(x_{1,n-1}) \\ \vdots & \vdots & & \vdots \\ \log(x_{m-1,0}) & \log(x_{m-1,1}) & \cdots & \log(x_{m-1,n-1}) \end{bmatrix}. \end{aligned}$$

对实数 x, 激活函数

$$\mathrm{silu}(x) = \frac{x}{1 + e^{-x}},$$

silu 的图像见图1. 函数 silu 作用到的向量和矩阵上时，它作用到每一个元素上。

均方层归一化（Root Mean Square Layer Normalization）函数

$$\mathrm{rnor}(X) = \begin{bmatrix} \frac{x_{11}}{\sigma_1} & \frac{x_{12}}{\sigma_1} & \cdots & \frac{x_{1n}}{\sigma_1} \\ \frac{x_{21}}{\sigma_2} & \frac{x_{22}}{\sigma_2} & \cdots & \frac{x_{2n}}{\sigma_2} \\ \vdots & \vdots & & \vdots \\ \frac{x_{m1}}{\sigma_m} & \frac{x_{m2}}{\sigma_m} & \cdots & \frac{x_{mn}}{\sigma_m} \end{bmatrix},$$

这里的 $\sigma_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n} x_{ij}^2}, i = 1, 2, \ldots, m$。

假设行向量 $\hat{\boldsymbol{x}} = (\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_{n-1})$，将行向量与矩阵相加定义为逐行相加，即

$$X + \hat{\boldsymbol{x}} = \begin{bmatrix} x_{00} + \hat{x}_0 & x_{01} + \hat{x}_1 & \cdots & x_{0,n-1} + \hat{x}_{n-1} \\ x_{11} + \hat{x}_0 & x_{11} + \hat{x}_1 & \cdots & x_{1,n-1} + \hat{x}_{n-1} \\ \vdots & \vdots & & \vdots \\ x_{m-1,0} + \hat{x}_0 & x_{m-1,1} + \hat{x}_1 & \cdots & x_{m-1,n-1} + \hat{x}_{n-1} \end{bmatrix}.$$

旋转位置编码 (Rotary Position Embeddings, RoPE)[1] 是大语言模型中的常用组件，其设计目标是“通过绝对位置编码的方式实现相对位置编码”。详细推导过程见设计者个人网站[2] 和 [5]。对任意给定的偶数 $n_6 \geq 2$，词碎序列长度 n_3。对 $\forall i = 0, 1, \ldots, n_3 - 1$，旋转矩阵

$$A_i = \begin{bmatrix} \cos i\bar{\theta}_0 & \sin i\bar{\theta}_0 & 0 & 0 & \cdots & 0 & 0 \\ -\sin i\bar{\theta}_0 & \cos i\bar{\theta}_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos i\bar{\theta}_2 & \sin i\bar{\theta}_2 & \cdots & 0 & 0 \\ 0 & 0 & -\sin i\bar{\theta}_2 & \cos i\bar{\theta}_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos i\bar{\theta}_{n_6-2} & \sin i\bar{\theta}_{n_6-2} \\ 0 & 0 & 0 & 0 & \cdots & -\sin i\bar{\theta}_{n_6-2} & \cos i\bar{\theta}_{n_6-2} \end{bmatrix},$$

显然矩阵 A_i 尺寸是 $n_6 \times n_6$，典型值是 128×128。对于弧度 $\bar{\theta}_t$，$t = 0, 2, 6, \ldots, n_6 - 2$，原始论文 [1] 使用固定值

$$\bar{\theta}_t = 10000^{-t/n_6}.$$

对任意实数行向量 $\boldsymbol{x} = (x_0, x_1, \ldots, x_{n_6-1})$ 和非负整数 i，定义旋转函数为

$$\mathrm{rope}(\boldsymbol{x}, i) = \boldsymbol{x} A_i. \tag{1}$$

利用矩阵 A_i 中元素的变化规律，将式 (1) 中的矩阵向量乘改写成向量乘，可以节约计算量

$$\mathrm{rope}(\boldsymbol{x}, i) = \boldsymbol{x} \otimes \boldsymbol{\xi}_1 + (-x_1, x_0, -x_3, x_2, \ldots, -x_{x_6-1}, x_{x_6-2}) \otimes \boldsymbol{\xi}_2,$$

这里的算符 $\otimes$ 表示向量按元素相乘，$\boldsymbol{\xi}_1 = (\cos i\bar{\theta}_0, \cos i\bar{\theta}_0, \cos i\bar{\theta}_2, \cos i\bar{\theta}_2, \ldots, \cos i\bar{\theta}_{n_6-2}, \cos i\bar{\theta}_{n_6-2})$，$\boldsymbol{\xi}_2 = (\sin i\bar{\theta}_0, \sin i\bar{\theta}_0, \sin i\bar{\theta}_2, \sin i\bar{\theta}_2, \ldots, \sin i\bar{\theta}_{n_6-2}, \sin i\bar{\theta}_{n_6-2})$。

对尺寸为 $n_3 \times n_6$ 的矩阵 X，逐行旋转

$$\mathrm{rope}(X, i) = \begin{bmatrix} \mathrm{rope}(x_{0:}, 0) \\ \mathrm{rope}(x_{1:}, 1) \\ \vdots \\ \mathrm{rope}(x_{n_3-1:}, n_3 - 1) \end{bmatrix} = (\mathrm{rope}(x_{0:}, 0); \mathrm{rope}(x_{1:}, 1); \ldots; \mathrm{rope}(x_{n_3-1:}, n_3 - 1)).$$

[2] https://kexue.fm/archives/8265

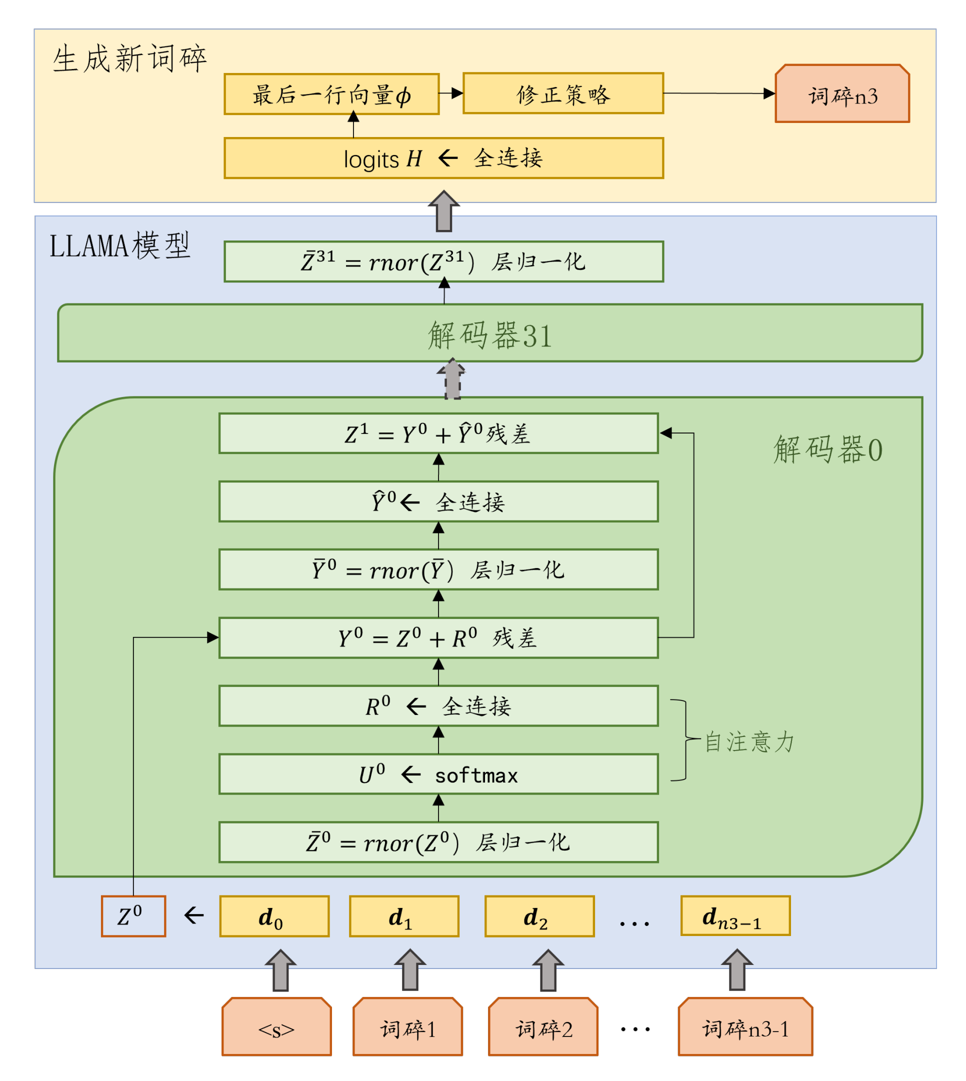


图 2: LLAMA-7B 模型全貌。词碎序列长度为 n_3，解码器层数为 32。

3 LLAMA 模型全貌

定义几个常数，并给出典型值。典型值是 Meta 公司预训练模型的一种参数配置，即 LLAMA-2-7B 的配置，其它参数配置见 LLAMA 源码网站[3]。n_1 为词表里的词碎数量，典型值 32000；n_2 为词碎嵌入宽度，典型值 4096；n_3 输入序列长度，正整数，进入模型前指定，在模型中保持不变；n_5 为自注意力头数，典型值 32；n_6 为单头宽度，等于 $\frac{n_2}{n_5}$，典型值 128；n_7 为全连接层宽度，典型值 11008；n_8 为解码器层数，典型值 32。需要注意，这几个常数的含义与 [4] 中同名常数的含义相同。

LLAMA 模型的全貌见图2。LLAMA 模型的输入是词碎序列，形式为

<s>␣词碎 1␣词碎 2␣词碎 3␣··· 词碎 n_3-2␣词碎 n_3-1

这里的 ␣ 是显式空格，用来分隔词碎。词碎 0 永远是 <s>，表示序列的开头。例如句子

`Who␣is␣the␣45th␣President␣of␣the␣United␣States?`

对应的词碎序列是

[3]https://huggingface.co/meta-llama

```
<s>␣Who␣is␣the␣4␣5␣th␣President␣of␣the␣United␣States␣?
```

词碎进入 LLAMA 之后，立即被转化为向量。相应地，词碎序列转化为矩阵 Z^0，Z^0 的每一个行向量对应一个词碎。接下来，矩阵 Z^0 被喂给第 0 个解码器，第 0 解码器输出矩阵 Z^1，矩阵 Z^1 随后被喂给第 1 个解码器。这样依次操作，第 n_8-1 个解码器的输出为 $\bar{Z}^{n_8}$，这也是 LLAMA 模型的输出。

4 制作输入序列

输入序列可以是任意指定的一段文本，然后转化为一个词碎序列，具体的转化方法有字对编码 Byte Pair Encoding、WordPiece 和 Unigram Language Model。词碎词典记为 $\mathcal{C}=\{c_0,c_1,\ldots,c_{n_1-1}\}$，词典中包含几个特殊的词碎，<unk>、<s>、</s>，含义分别为未定义、序列开头、序列结尾。对中文来说，词碎是单个字、单个标点符号、单个字对应的字节。例如" 你"、" 好" 对应的词碎是它们自身" 你"、" 好"，而" 啊" 被拆分成 3 个词碎 <0xE5>␣<0x95>␣<0x8A>。对英文来说，词碎是组成单词的片段，任何一个单词都可以分割成若干词碎，例如`unaffable`能分割成`una␣ff␣able`。

将输入文本中的中文、英文全部转化为词碎，就得到词碎形式的输入序列，此后提及的输入序列均指词碎形式的输入序列。

词典的中每个词碎 c_i 都嵌入到一个行向量 $\boldsymbol{d}_i$，$\boldsymbol{d}_i$ 的尺寸为 $1\times n_2$，尺寸典型值为 1×4096。将所有的行向量 $\boldsymbol{d}_i$ 按顺序排列起来，组成矩阵 $D=(\boldsymbol{d}_1;\boldsymbol{d}_2;\ldots;\boldsymbol{d}_{n_1})$，尺寸为 $n_1\times n_2$，尺寸典型值 32000×4096。输入序列序列记为 $\boldsymbol{\tau}$，形式为

$$\boldsymbol{\tau}=\tau_0\tau_1\ldots\tau_{n_3-1},$$

这里的 $\tau_i\in\mathcal{C}, i=0,1,\ldots,n_3-1$。输入序列 $\boldsymbol{\tau}$ 中词碎的位置编号记为 $\boldsymbol{t}=(t_0,t_1,\ldots,t_{n_3-1})$，在词典 $\mathcal{C}$ 中的编号记为 $\hat{\boldsymbol{t}}=(\hat{t}_0,\hat{t}_1,\ldots,\hat{t}_{n_3-1})$，显然 $t_i\in\{0,1,\ldots,n_3-1\}$，$\hat{t}_i\in\{0,1,\ldots,n_1-1\}$。

输入序列

```
<s>␣Who␣is␣the␣4␣5␣th␣President␣of␣the␣United␣States␣?
```

的位置编码为

$$\boldsymbol{t}=(0,1,2,3,4,5,6,7,8,9,10,11,12,13),$$

在词典 $\mathcal{C}$ 中的编号是

$$\hat{\boldsymbol{t}}=(1,11644,338,278,29871,29946,29955,386,7178,310,278,3303,3900,29973).$$

5 遮挡矩阵

$$M=\begin{bmatrix} 0 & -\infty & -\infty & \cdots & -\infty \\ 0 & 0 & -\infty & \cdots & -\infty \\ 0 & 0 & 0 & \cdots & -\infty \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

矩阵 M 的尺寸是 $n_3\times n_3$，典型值 128×128。主对角线以及下三角元素值为 0，上三角元素值为 $-\infty$。在实际计算时，$-\infty$ 的取值是 M 所属数据类型所能表示的最小值。例如`torch.float32`的最小值是 -3.40282×10^{38}，`torch.float16`的最小值是 -65504。

6 解码器

输入 LLAMA 模型的样本是词碎序列，不能直接进行矩阵运算、向量运算，需要先转换成矩阵形式，即矩阵 Z^0。这个转换工作在第 0 个解码器前完成。给定输入序列 $\boldsymbol{\tau}=\tau_0\tau_1\ldots\tau_{n_3-1}$。对 $i=0,1,\ldots,n_3-1$，将 D 的第 $\hat{t}_i$ 取出来，放在矩阵 Z^0 的第 t_i 行。Z^0 的尺寸为 $n_3\times n_2$，典型值为 $n_3\times 4096$。

本节里的各个子层均在第 0 个解码器，不再每次说明。

6.1 层归一化子层

$$\bar{Z}^0=\mathrm{rnor}(Z^0),$$

矩阵 $\bar{Z}^0$ 的尺寸为 $n_3\times n_2$，典型值为 $n_3\times 4096$。

6.2 自注意力子层

引入“查”权重矩阵 W^{01}，尺寸 $n_2\times n_2$，典型值 4096×4096，上标 0 对应解码器的编号。将 $(W^{01})^T$ 简记为 W^{01T}。“查”矩阵

$$Q^0=\bar{Z}^0W^{01T},$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。引入“键”权重矩阵 W^{02}，尺寸 $n_2\times n_2$，典型值 4096×4096，上标 0 对应解码器的编号。“键”矩阵

$$K^0=\bar{Z}^0W^{02T},$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。引入“值”权重矩阵 W^{03}，尺寸 $n_2\times n_2$，典型值 4096×4096，上标 0 对应解码器的编号。“值”矩阵

$$V^0=\bar{Z}^0W^{03T},$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。

将“查”矩阵 Q^0 按列平均分块，每个小块记为矩阵 $Q^{0,i}$，$i=0,1,2,\ldots,n_5-1$，即

$$Q^0=[Q^{0,0},Q^{0,1},\cdots,Q^{0,n_5-1}].$$

$Q^{0,i}$ 的尺寸是 $n_3\times n_6$，典型值 $n_3\times 128$。将小块矩阵旋转得到 $\bar{Q}^{0,i}=\mathrm{rope}(Q^{0,i})$，尺寸与 $Q^{0,i}$ 相同，为 $n_3\times n_6$，典型值 $n_3\times 128$。

将“键”矩阵 K^0 按列平均分块，每个小块记为矩阵 $K^{0,i}$，$i=0,1,2,\ldots,n_5-1$，即

$$K^0=[K^{0,0},K^{0,1},\cdots,K^{0,n_5-1}].$$

$K^{0,i}$ 的尺寸是 $n_3\times n_6$，典型值 $n_3\times 128$。将小块矩阵旋转得到 $\bar{K}^{0,i}=\mathrm{rope}(K^{0,i})$，尺寸与 $K^{0,i}$ 相同，为 $n_3\times n_6$，典型值 $n_3\times 128$。

将“值”矩阵 V^0 按列平均分块，每个小块记为矩阵 $V^{0,i}$，$i=0,1,2,\ldots,n_5-1$，即

$$V^0=[V^{0,0},V^{0,1},\cdots,V^{0,n_5-1}].$$

$V^{0,i}$ 的尺寸是 $n_3\times n_6$，典型值 $n_3\times 128$。记

$$U^{0i}=\text{smax}\left(\frac{\bar{Q}^{0,i}(\bar{K}^{0i})^T}{\sqrt{n_6}}+M\right)V^{0,i},$$

尺寸是 $n_3\times n_6$，典型值 $n_3\times 128$。将 n_5 个小矩阵按行拼接成大矩阵

$$U^0=[U^{0,0},U^{0,1},\cdots,U^{0,n_5-1}],$$

尺寸为 $n_3\times n_2$，典型值 $n_3\times 4096$。

引入“出”权重矩阵 W^{04}，尺寸 $n_2\times n_2$，典型值 4096×4096。“出”矩阵为

$$R^0=U^0W^{04T},$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。

6.3 残差子层

$$Y^0=Z^0+R^0,$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。

6.4 全连接子层

令

$$\bar{Y}^0=\text{rnor}(Y^0),$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。引入“门”权重矩阵 W^{05}，尺寸 $n_7\times n_2$，典型值 11008×4096；“下”权重矩阵 W^{06}，尺寸 $n_2\times n_7$，典型值 4096×11008；“上”权重矩阵 W^{07}，尺寸 $n_7\times n_2$，典型值 11008×4096。令

$$\hat{Y}^0=\left(\text{silu}(\bar{Y}^0W^{05T})\otimes(\bar{Y}^0W^{07T})\right)W^{06T},$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。算符 $\otimes$ 表示两个矩阵相同位置的元素相乘。解码器 0 的输出为

$$Z^1=Y^0+\hat{Y}^0,$$

尺寸 $n_3\times n_2$，典型值 $n_3\times 4096$。

6.5 解码器堆叠

第 0 个解码器的输入是矩阵是 Z^0，输出矩阵是 Z^1。每个解码器内部的计算过程都一样，第 1 个解码器的输入矩阵是 Z^1，输出矩阵是 Z^2。依次类推，第 n_8-1 个解码器的输入矩阵是 Z^{n_8-1}，输出矩阵是 Z^{n_8}。对 $j=0,1,\ldots,n_8-1$，矩阵 Z^j 的尺寸是 $n_3\times n_2$，典型值是 $n_3\times 4096$。

6.6 模型输出

$\bar{Z}^{n_8}=\text{rnor}(Z^{n_8})$ 是 LLAMA 模型的输出，尺寸是 $n_3\times n_2$，典型值是 $n_3\times 4096$。

7 生成下一个词碎

引入权重矩阵 $\bar{W}^8$，尺寸 $n_1 \times n_2$，典型值 32000×4096，对分数 (logit) 矩阵

$$H = \bar{Z}^{n_8} \bar{W}^{8T},$$

尺寸是 $n_3 \times n_1$，尺寸典型值是 $n_3 \times 32000$。取出矩阵 H 的最后一行，即第 n_3-1 行，记为向量 $\boldsymbol{\phi}$，向量长度 n_1，典型值 32000。$\boldsymbol{\phi}$ 称为“下对分”，即下一个词碎的对分数。记 $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_{n_1-1})$。

为了引入随机性、减少重复性等目的，生成下一个词碎前，还可对对分数进行修正。

7.1 分数修正策略：重复惩罚

惩罚系数是一个任意指定的超参数，$\alpha_1 > 0$。使用 α_1 修改 $\boldsymbol{\phi}$ 的元素值，具体做法是，对 $\hat{\boldsymbol{t}} = (\hat{t}_0, \hat{t}_1, \ldots, \hat{t}_{n_3-1})$ 指定位置的元素放缩，小于 0 的放大 α_1 倍，大于等于 0 的缩小 α_1 倍，即对 $i \in \{\hat{t}_0, \hat{t}_1, \ldots, \hat{t}_{n_3-1}\}$，令

$$\bar{\phi}_i = \begin{cases} \alpha_1 \phi_i, & \text{如果 } \phi_i < 0, \\ \frac{\phi_i}{\alpha_1}, & \text{如果 } \phi_i \geq 0. \end{cases}$$

然后用 $\bar{\phi}_i$ 替换 $\boldsymbol{\phi}$ 中的第 i 个元素。

7.2 分数修正策略：温度

温度 (temperature) 是一个任意指定的超参数，$\alpha_2 > 0$。使用温度修改 $\boldsymbol{\phi}$ 的元素值，即 $\bar{\boldsymbol{\phi}} = \boldsymbol{\phi}/\alpha_2$。为了方便叙述，修正后的分数仍然记为 $\boldsymbol{\phi}$。

7.3 概率采样

令 $\hat{\boldsymbol{\phi}} = \mathrm{smax}(\boldsymbol{\phi})$，向量长度 n_1，典型值 32000。从整数 $\{0, 1, 2, \ldots, n_1 - 1\}$ 中随机选出一个数，整数 i 被选中的概率为 $\hat{\phi}_i$。被选出来的编号记为 $\hat{t}_{n_3}$，对应的词碎记为 τ_{n_3}，将 τ_{n_3} 追加到序列输入序列 $\boldsymbol{\tau}$ 的尾部。如果 τ_{n_3} 是 </s> 或者序列 $\boldsymbol{\tau}$ 达到指定长度，那么生成过程结束。否则，继续生成更多词碎。

8 微调训练

任意给定一个句子，用碎词机 (tokenizer) 将这个句子切成词碎序列 $\boldsymbol{\tau} = \tau_0\tau_1 \ldots \tau_{n_3-1}$，长度为 n_3，这里的 $\tau_i \in \mathcal{C}, i = 0, 1, \ldots, n_3 - 1$。$\boldsymbol{\tau}$ 中词碎在词典 $\mathcal{C}$ 中的编号为 $\hat{\boldsymbol{t}} = (\hat{t}_0, \hat{t}_1, \ldots, \hat{t}_{n_3-1})$，例如句子

`Who␣is␣the␣45th␣President␣of␣the␣United␣States?␣Donald␣Trump.`

对应的词碎序列为

`<s>␣Who␣is␣the␣4␣5␣th␣President␣of␣the␣United␣States␣?␣Donald␣Trump␣.`

记为 $\boldsymbol{\tau}^1$。序列长度为 17，即此时 $n_3 = 17$，序列 $\boldsymbol{\tau}^1$ 在词典 $\mathcal{C}$ 中的编号是

$$\hat{\boldsymbol{t}}^1 = (1, 11644, 338, 278, 29871, 29946, 29955, 386, 7178, 310, 278, 3303, 3900, 29973, 18935, 27504, 29889).$$

将序列 $\boldsymbol{\tau}$ 输入模型，计算得到对分数矩阵 H，尺寸是 $n_3\times n_1$，尺寸典型值是 $n_3\times 32000$。将 H 的最后一行去掉，得到一个新的对分数矩阵

$$\bar{H}=\begin{bmatrix}\bar{h}_{00} & \bar{h}_{01} & \cdots & \bar{h}_{0,n_1-1}\\ \bar{h}_{10} & \bar{h}_{11} & \cdots & \bar{h}_{1,n_1-1}\\ \vdots & \vdots & & \vdots\\ \bar{h}_{n_3-2,0} & \bar{h}_{n_3-2,1} & \cdots & \bar{h}_{n_3-2,n_1-1}\end{bmatrix}=\begin{bmatrix}h_{00} & h_{01} & \cdots & h_{0,n_1-1}\\ h_{10} & h_{11} & \cdots & h_{1,n_1-1}\\ \vdots & \vdots & & \vdots\\ h_{n_3-2,0} & h_{n_3-2,1} & \cdots & h_{n_3-2,n_1-1}\end{bmatrix},$$

尺寸是 $n_3-1\times n_1$，尺寸典型值是 $n_3-1\times 32000$。当输入序列是 $\boldsymbol{\tau}^1$ 时，$\bar{H}$ 的尺寸是 16×32000。

8.1 微调格式 1

截取 $\hat{\boldsymbol{t}}$ 尾部的 n_3-1 元素，记为 $\bar{\boldsymbol{t}}$，即

$$\bar{\boldsymbol{t}}=(\bar{t}_0,\bar{t}_1,\bar{t}_2,\ldots,\bar{t}_{n_3-2})=(\hat{t}_1,\hat{t}_2,\hat{t}_3,\ldots,\hat{t}_{n_3-1}),$$

长度是 n_3-1。当输入序列是 $\boldsymbol{\tau}^1$ 时，

$$\bar{\boldsymbol{t}}^1=(11644,338,278,29871,29946,29955,386,7178,310,278,3303,3900,29973,18935,27504,29889),$$

长度为 16。

8.2 微调格式 2

对问答型任务，可以将问句词碎对应的位置编号置为 <unk>，在 LLAMA 中值为-100。例如，当输入序列是 $\boldsymbol{\tau}^1$ 时，

$$\bar{\boldsymbol{t}}^1=(-100,-100,-100,-100,-100,-100,-100,-100,-100,-100,-100,-100,-100,18935,27504,29889),$$

长度为 16，尾部 3 个不等于-100 的值，对应着答句

`Donald␣Trump␣.`

8.3 交叉熵

令

$$\tilde{H}=\mathrm{smax}(H),$$

尺寸是 $n_3-2\times n_1$，尺寸典型值是 $n_3-2\times 32000$。当输入序列是 $\boldsymbol{\tau}^1$ 时，$\bar{H}$ 的尺寸是 16×32000。损失函数为

$$L=-\sum_{\substack{i=0\\ \bar{t}_i\neq -100}}^{n_3-2}\log(\tilde{h}_{i,\bar{t}_i}).$$

$\tilde{H}$ 的第 i 行向量是前 $i+1$ 个词碎生成的第 $i+2$ 个词碎的概率分布，$\bar{t}_i$ 是真实输入序列中 $i+2$ 个词碎的编号。训练的目标是使 $\bar{t}_i$ 无限接近于 0，即模型生成的词碎与输入序列中词碎相同。

9 术语对应关系

本文中的数学公式全部提取自 LLAMA 模型源码，为方便理解，列出源码中对象的含义、对应的数学符号、典型值。

vocab_size：词碎数量，n_1，32000。

hidden_size：词碎嵌入宽度，n_2，4096。

seq_len，词碎序列长度，n_3，每迭代一步加 1，LLAMA-7B 允许的最大值 2049，LLAMA-2-7B 允许的最大值 4096。

num_attention_heads：自注意力头数，n_5，32。

head_dim：单头宽度，n_6，128。

intermediate_size：全连接层宽度，n_7，11008。

num_hidden_layers：解码器层数，n_8，32。

q_proj：查权重矩阵，W^{01}，尺寸 $n_2 \times n_2$，典型值 4096×4096。

k_proj：键权重矩阵，W^{02}，尺寸 $n_2 \times n_2$，典型值 4096×4096。

v_proj：值权重矩阵，W^{03}，尺寸 $n_2 \times n_2$，典型值 4096×4096。

o_proj：出权重矩阵，W^{04}，尺寸 $n_2 \times n_2$，典型值 4096×4096。

gate_proj：门权重矩阵，W^{05}，尺寸 $n_7 \times n_2$，典型值 11008×4096。

down_proj：下权重矩阵，W^{06}，尺寸 $n_2 \times n_7$，典型值 4096×11008。

up_proj：上权重矩阵，W^{07}，尺寸 $n_7 \times n_2$，典型值 11008×4096。

lm_head：W^{08}，尺寸 $n_1 \times n_2$，典型值 32000×4096。

mlp：全连接，包含 W^{05}、W^{06}、W^{07}。

self_attn：自注意力，包含 W^{01}、W^{02}、W^{03}、W^{04}。

对 baichuan-7B 来说，W_pack是q_proj、k_proj、v_proj的按列拼接形成的 3 倍大矩阵，即 $[W^{01}, W^{02}, W^{03}]$。

10 参数量

需要训练的参数是 $W^{j1} \sim W^{j7}$ 和 W^{08}，$j = 0, 1, 2, \ldots, n_8 - 1$，从而参数数量为 $n_8(4n_2^2 + 3n_2n_7) + n_1n_2 = 4n_2^2n_8 + 3n_2n_7n_8 + n_1n_2$。

对 LLAMA-7B，参数数量为 66 0707 7376。对 LLAMA-13B，$n_1 = 32000$，$n_2 = 5120$，$n_5 = 40$，$n_6 = 128$，$n_7 = 13824$，$n_8 = 40$，参数数量为 128 5160 9600。对 Baichuan-7B，$n_1 = 64000$，$n_2 \sim n_8$ 的取值与 LLAMA-7B 相同，参数数量为 67 3814 9376。对 Baichuan-13B，$n_1 = 64000$，$n_2 \sim n_8$ 的取值与 LLAMA-13B 相同，参数数量为 130 1544 9600。

参考文献

[1] Jianlin Su. *RoFormer: Transformer with Rotary Position Embeddings - ZhuiyiAI*. Tech. rep. 2021. URL: https://github.com/ZhuiyiTechnology/roformer.

[2] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[3] Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models". In: (2023). arXiv: 2302.13971 [cs.CL].

[4] 何沧平; 许涛; "BERT 模型的数学形式". In: (). DOI: 10.12074/202110.00071.

[5] 何沧平; 许涛; “大语言模型旋转位置编码的简易推导”. In: (). DOI: 10.12074/202307.00071V2.